

Express Mail EK873465687US
Mailed July 9, 2001

SYSTEM AND METHOD FOR IMPLEMENTING A ROBOT PROOF WEB SITE

Field of the Invention

The present invention relates to the Internet and more particularly applies to those of the World Wide Web (WWW) sites that, while welcoming human beings, want to exclude robots from visiting and gathering information from them.

Background of the Invention

WWW robots, also called Web Wanderers, Web Crawlers or Web Spiders, and often just referred to as bots (bot is short for robot), are programs devised to automatically traverse the hypertext structure of the Web. Such bots, having retrieved a document, can also recursively retrieve all the linked pages referenced in the document. This is especially the case of numerous search engines and their robots which roam the World Wide Web finding and indexing content to add to their databases. Although most robots provide a valuable service, concern has developed amongst Web site administrators about exactly how much

of their precious server time and bandwidth is being used to service requests from these engines.

While the majority of robots are well designed, are professionally operated and cause no problems, there are occasions where robots visiting Web servers are not welcome because of the way robots behave. Some may swamp servers with rapid-fire requests, or retrieve the same files repeatedly. If done intentionally this is a form of Denial of Service (DoS) attack, although this is more often just the result of a poor or defective robot design. In other situations robots traverse parts of WWW servers that are not suitable for being searched e.g., contain duplicated or temporary information, include large documents or e.g., CGI scripts (CGI is a standard for running external programs from a World-Wide Web HTTP server). In this latter case and in similar situations, when accessed and executed, scripts tend to consume significant server resources in generating dynamic pages and thus, slow down the system.

In recognition of these problems many Web robots offer facilities for Web site administrators and content providers to limit what the robot is allowed to do. Two mechanisms are provided. One is referred to as the 'Robots Exclusion Protocol', even though it is not really an enforced protocol, but was a working draft document discussed as an Internet-Draft by the Internet Engineering Task Force (IETF) in 1996 under the title 'A Method for Web Robots Control'. According to this document, a Web site administrator can indicate which parts of the site should not be visited by a robot. This is accomplished by providing a specially formatted file, in <http://.../robots.txt>. The second

mechanism assumes that a Web author can indicate whether a page may or may not be indexed, or analyzed for links, through the use of a special Hyper Text Markup Language (HTML) META tag i.e., a 'Robots META tag'. However, both of these mechanisms rely on cooperation from the robots, and are not even guaranteed to work for every robot. Moreover, as already suggested here above relative to DoS attacks, some of these robots may not be so friendly. They could be run e.g., with the malicious intent of attacking a Web site (then, they just ignore the robots.txt file and the robots meta tags) so the site becomes overloaded and starts refusing to serve legitimate users i.e., the human beings trying to make normal use of the site.

Also, although the information made available on a site may not be confidential, an administrator may want to prevent the unlimited dissemination of it that would otherwise result from the indexing and referencing activities of all sorts of robots. The standard way of achieving this is to protect a Web site through some form of authentication, of which the more common method is to manage a list of registered users having a password so as they have to sign on upon accessing the site. The obvious drawback of this is that administrators must manage and update a closed list of users. This requires a registration step for a first consultation of a site and also assumes that users will remember their passwords in subsequent consultations. This may not be at all what the administrator wanted to achieve, and may even be counterproductive, since it will certainly discourage some individuals who are willing to browse a site to go further if they are requested to register.

Object of the Invention

Thus, it is a broad object of the invention to prevent Web site contents from being investigated by robots.

It is a further object of the invention to encourage human beings, attempting to access a robot protected Web site, to proceed further by not imposing a registration at first access and a log on procedure at each subsequent access.

It is still another object of the invention to avoid reliance on a robot's cooperation for barring the robot's access to contents of Web sites.

Further objects, features and advantages of the present invention will become apparent to the ones skilled in the art upon examination of the following description in reference to the accompanying drawings. It is intended that any additional advantages be incorporated herein.

Summary of the Invention

A method and a system for preventing robots from browsing a Web site beyond a welcome page are described. On receiving an initial request from an undefined originator, the Web site responds to the request with a welcome page including a challenge. Then, on receiving a further request from the undefined originator, the Web site can check whether the challenge is fulfilled or not. If fulfilled, the undefined originator is assumed to be a human being and the site keeps

processing the further request and all subsequent ones, if any. However, if the challenge is not fulfilled, the undefined originator is assumed to be a robot, in which case all requests from that originator are not further processed.

The invention prevents Web site contents from being investigated by robots without requiring end users to register or the site administrator to have to manage an access list of authorized users.

Brief Description of the Drawings

Figure 1 is an exemplary welcome page per the invention.

Figure 2 shows the corresponding HTML code.

Figure 3 shows the steps of the method of the invention.

Figure 4 shows the further steps of the method when access to a Web site per the invention is denied, while a timer is on, for requests carrying a logged IP address.

Figure 5 are other exemplary welcome pages with challenges.

Detailed Description of the Preferred Embodiment

Figure 1 illustrates the method according to the invention to prohibit robots from accessing a Web site beyond its welcome page. An exemplary welcome page as seen by an individual accessing a Web site e.g., at following URL [100] (Uniform Resource Locator) 'http://www.anysite.com/welcome.html' is shown. Accessing a Web site can be done with any available Web browser e.g., Netscape browser [110] from Netscape Communications Corporation, 501 E. Middlefield Road, Mountain View, CA 94043, USA can be used. Then, according to a first method to prohibit robot's access, the welcome page implements a dummy challenge that can be simply answered by a human being, but which a robot should certainly fail.

Among various possibilities, Figure 1 illustrates a typical challenge according to the invention. The welcome page thus shows an image [120] including, in this particular example, a few geometric forms that can be unambiguously referred to by a single word or expression in a language that the individual accessing the Web is assumed to be capable of reading. Then, associated with the image which, as depicted here, includes a square, a circle, a cross and a triangle, the reader is prompted [130] to click e.g., on the cross [121]. A human being, desiring to go on and visit the site, will do it naturally while a robot will do nothing, or will do it wrongly simply because it just does not

understand the graphic symbols or which of them may be a cross. Hence, this invention easily allows discrimination of a human being from a robot on the basis of their respective levels of abstract reasoning, which is naturally high, or very high, for a human being while a robot is totally lacking this capacity. This allows achievement of the objectives of the invention to prevent robots from browsing the site beyond the welcome page, while neither imposing to the people accessing it the burden of having to register and to log on, nor requiring the administrators of the site to manage a list of legitimate users.

To make the site even more resistant to browsing by a "smart" robot that could be tailored to adapt to a given challenge or set of predictable such challenges, prompting can be made random so that every time someone accesses the site, the challenge is different. For example, the cross can be moved to a different position on the image map so that the coordinates returned, when clicking on it, are different. Or, the prompting request can change in requesting e.g., to click on triangle instead.

Figure 2 shows the source HTML (Hyper Text Markup Language) code [140] for this example which uses a server side map i.e., ismap [142], included in an anchor tag created with the `<a..> ..` construct form [149] of the HTML language. Thus, when the user clicks on the cross [121], the browser sends a request back to the server URL (/cgi-bin/challenge.exe) [141] including the Xand Y coordinates of the user's click contained in an ismap

[142] so that the server can check whether the click coordinates indeed matche the cross position. The anchor tag also carries an identification field i.e., id=XD34F739 [143] which is useful to correlate the answer, including the click coordinates, with the current challenge when this latter changes from one user to another as explained above.

Figure 3 depicts the steps of the method according to the invention when an originator of an initial request to access a Web site is responded to with a challenge. Upon receiving this initial request, the [300] Web site server responds with the challenge [310]. This is done through the establishment of a TCP connection with the originator (the reliable transport protocol of the Internet TCP/IP suite of protocols used by the Web). The response is in the form of a Web page including a challenge e.g., of the kind discussed in Figure 1. Then, having gotten the server's response, the originator proceeds with a new transaction containing his click response directed towards the Web site [320]. On receiving the new transaction, the Web site server checks whether the challenge is fulfilled or not [330]. If the response matches the challenge prompt [340], then the site server assumes that the originator is a human being and permits further access. However, if the Web server finds that challenge is not properly answered, it must assume that the originator is a robot [350]. As a consequence, it stops processing all current and further requests, if any [351], which includes dropping the TCP connection or redirecting it to another site [352]. Also, the IP source address [361] may be remembered and a timer started [362] so that access to the site may be temporarily barred, from

that IP source address, as explained in Figure 4.

Figure 4 shows the case where the IP address of an assumed robot originator is remembered. Then, the server may be implemented to temporarily drop or redirect immediately all requests issued with this particular source IP address, because all subsequent access requests thereafter are assumed to come from a robot (although this might not always be true since a robot may be behind a proxy or firewall performing a network address translation of all the IP source addresses it has to forward). Because the IP source address of the request was logged and a timer started as explained in Figure 3, each time a new request is received [410] the server first checks for whether the same logged source IP address is found [420]. If not the same address, access may proceed normally [450]. If the address is the same, however, a timer is checked [430]. If the timer interval has elapsed, the logged IP address is reset [440] and new request is normally handled [450], i.e. a new challenge is issued. However, if timer has not elapsed, the TCP connection is dropped or redirected [460] before the server resumes handling of newly received requests [410].

Figure 5 illustrates other exemplary challenge examples that are easily answered by human beings.

Figure 5-a takes the form of a quiz [510] which could be made as simple as shown [500] or as sophisticated as necessary to defeat elaborated or "smart" robots or, alternatively, to adapt to a particular end-user population sharing a common type of skill.

Figure 5-b is another alternative that combines images [520]

and text [530] in an even more abstract way where the answer is suggested by the text, so it is even better adapted to discriminate a human being from a robot. However, it is worth mentioning here that such a challenge is culture-dependent and could serve as well to discriminate human beings on the basis of their social or ethnic origins.

CONFIDENTIAL